

Anonymized Local Privacy

Joshua Joy¹ and Mario Gerla¹

¹UCLA

{jjoy,gerla}@cs.ucla.edu

Abstract—In this paper, we introduce the family of Anonymized Local Privacy mechanisms. These mechanisms have an output space of multiple values (e.g., “Yes”, “No”, or “⊥” (not participating)) and leverage the law of large numbers to generate linear noise in the number of data owners to protect privacy both before and after aggregation yet preserve accuracy.

We describe the suitability in a distributed on-demand network and evaluate over a real dataset as we scale the population.

I. INTRODUCTION

Personal mobile information is being continuously collected and analyzed with minimal regard to privacy. As we transition from small mobile personal devices to large-scale sensor collecting self-driving vehicles the needs of privacy increase.

Differential privacy has emerged as the gold standard for privacy protection. Differential privacy essentially states that whether or not a single data owner decides to participate in data collection, the final aggregate information will be perturbed only by a negligible amount. That is, the aggregate information released gives no hints to the adversary about a particular data owner. However, differential private techniques do not add noise linear in the number of data owners to protect. Techniques, such as the Laplace mechanism, add noise calibrated to the sensitivity of the query output, rather than linear in the number of data owners to protect, in order to preserve accuracy [7].

We introduce the family of Anonymized Local Privacy mechanisms and present constructions with better accuracy than randomized response.

Randomized response has been shown to be optimal in the local privacy setting [4]. However, in order to preserve accuracy with the randomized response mechanism, privacy must be sacrificed as the data owners must respond truthfully too frequently. For example, a data owner should respond truthfully more than 80% of the time to have decent accuracy which greatly minimizes any privacy gains [15], [14]. The reason is due to the high variance from the coin tosses [20]. As more aggressive sampling is performed, the variance quickly increases making it difficult to perform accurate estimation of the underlying distribution.

As a result of the accuracy problem, there have been various privacy-preserving systems which focus on the heavy-hitters only [9], [2]. These techniques ensure privacy only for large populations and can only detect or estimate the most frequently occurring distributions, rather than smaller or less frequently occurring populations.

Our contribution is the ability to maintain strong privacy while maintaining the fidelity of the data. The output space of

Anonymized Local Privacy mechanisms is three values “Yes”, “No”, or “⊥” (not participating) as opposed to solely two values “Yes” or “No”. Three output values allows for robust estimation, as we show in Section §??.

We evaluate the accuracy of our privacy-preserving approach utilizing a vehicular crowdsourcing scenario comprising of approximately 50,000 records. In this dataset, each vehicle reports its location utilizing the California Transportation Dataset from magnetic pavement sensors (see Section §??).

II. RELATED WORK

Differential privacy [5], [7], [6], [8] has been proposed as a mechanism to privately share data such that anything that can be learned if a particular data owner is included in the database can also be learned if the particular data owner is not included in the database. To achieve this privacy guarantee, differential privacy mandates that only a sublinear number of queries have access to the database and that noise proportional to the global sensitivity of the counting query is added (independent of the number of data owners).

Distributional privacy [1] is a privacy mechanism which says that the released aggregate information only reveals the underlying ground truth distribution and nothing else. This protects individual data owners and is strictly stronger than differential privacy. However, it is computationally inefficient though can work over a large class of queries known as Vapnik-Chervonenkis (VC) dimension.

Zero-knowledge privacy [12] is a cryptographically influenced privacy definition that is strictly stronger than differential privacy. Crowd-blending privacy [11] is weaker than differential privacy; however, with a pre-sampling step, satisfies both differential privacy and zero-knowledge privacy. However, these mechanisms do not add noise linear in the number of data owners and rely on aggressive sampling, which negatively impact the accuracy estimations.

The randomized response based mechanisms [22], [10], [13], [21] satisfies the differential privacy mechanism as well as stronger mechanisms such as zero-knowledge privacy. However, the accuracy of the randomized response mechanism quickly degrades unless the coin toss values are configured to large values (e.g., greater than 80%).

III. PRELIMINARIES

Differential Privacy. Differential privacy has become the *gold standard* privacy mechanism which ensures that the output of

a sanitization mechanism does not violate the privacy of any individual inputs.

Definition 1 ([5], [7]). (ϵ -Differential Privacy). A privacy mechanism $\text{San}()$ provides ϵ -differential privacy if, for all datasets D_1 and D_2 differing on at most one record (i.e., the Hamming distance $H()$ is $H(D_1, D_2) \leq 1$), and for all outputs $O \subseteq \text{Range}(\text{San}())$:

$$\sup_{D_1, D_2} \frac{\Pr[\text{San}(D_1) \in O]}{\Pr[\text{San}(D_2) \in O]} \leq \exp(\epsilon) \quad (1)$$

That is, the probability that a privacy mechanism San produces a given output is almost independent of the presence or absence of any individual record in the dataset. The closer the distributions are (i.e., smaller ϵ), the stronger the privacy guarantees become and vice versa.

Private Write. More generally, we assume some class of private information storage [19] mechanisms are utilized by the data owner to cryptographically protect their writes to cloud services.

IV. ANONYMIZED LOCAL PRIVACY

First, we define the structure of an anonymized local private mechanism. We then illustrate various mechanisms that satisfy Anonymized Local Privacy. Finally, we provide the mechanism for preserving accuracy in the Anonymized Local Privacy model.

A. Structure

An anonymized local privacy mechanism answers “Yes”, “No”, or \perp (not participating). For our purposes, we use the notation of the *Yes* population as the ground truth and the remaining data owners are the *No* population. Each population should blend with each other such that the aggregate information that is released is unable to be used to increase the confidence or inference of an adversary that is trying to determine the value of a specific data owner.

Data owners are aggressively sampled (e.g., 5%). To overcome the estimation error due to the large variance, the estimation of the noisy “Yes” counts and sampled counts are combined to offset each other and effectively cancel the noise, allowing for the aggressive sampling.

B. Sampling

Sampling whereby a centralized aggregator randomly discards responses has been previously formulated as a mechanism to amplify privacy [3], [18], [16], [17], [11]. The intuition is that when sampling approximates the original aggregate information, an attacker is unable to distinguish when sampling is performed and which data owners are sampled. These privacy mechanisms range from sampling without a sanitization mechanism, sampling to amplify a differentially private mechanism, sampling that tolerates a bias, and even sampling a weaker privacy notion such as k -anonymity to amplify the privacy guarantees.

However, sampling alone has several issues. First, data owners that answer “Yes” do not have the protection of strong plausible deniability as they never respond “No” or are “forced” to respond “Yes” (e.g., via coin tosses). Data owners that answer “No” do not provide privacy protection as they never answer “Yes”. Second, as we increase the sampling rate the variance will increase rapidly, thus weakening accuracy. Finally, the privacy strength of the population does not increase as the population increases. The *Yes* population is fixed (e.g., those at a particular location) and we can only increase the *No* population. The *No* population should also contribute noise by answering “Yes” in order to strengthen privacy.

C. Sampling and Noise

We could leverage the *No* population by use the same sampling rate though for the *No* population have a portion respond “Yes”. To perform the final estimation we simply subtract the estimated added noise.

Sampling and Noise Response. Each data owner privatizes their actual value *Value* by performing the following Bernoulli trial. Let π_s be either the sampling probability for the *Yes* population as $\pi_{s_{Yes}}$ or for the *No* population as $\pi_{s_{No}}$.

$$\text{Privatized Value} = \begin{cases} \perp & \text{with probability } 1 - \pi_s \\ \text{Value} & \text{with probability } \pi_s \end{cases} \quad (2)$$

That is, a percentage of the *Yes* population responds “Yes” and a percentage of the *No* population responds “Yes” (providing noise). However, the *Yes* data owners do not answer “No” and also do not have plausible deniability (that is being forced via coin toss to respond “Yes”).

D. Sampling and Plausible Deniability

We would like to have a percentage of each population respond opposite of their actual value, provide plausible deniability, and have outputs from the space of “Yes”, “No”, and \perp (not participating) in order for the data owners to blend with each other.

To achieve plausible deniability via coin tosses we have a small percentage of the “Yes” population be “forced” to respond “Yes”. The other output values follow from the sampling and noise scenario.

$$\text{Privatized Value} = \begin{cases} \perp & \text{with probability } 1 - \pi_{s_{Yes}} \\ 1 & \text{with probability } \pi_{s_{Yes}} \times (\pi_1 + (1 - \pi_1) \times \pi_2) \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

$$\text{Privatized Value} = \begin{cases} \perp & \text{with probability } 1 - \pi_{s_{No}} \\ 1 & \text{with probability } \pi_{s_{No}} \times ((1 - \pi_1) \times \pi_2) \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

A benefit of plausible deniability is that the estimation of the population will provides privacy protection via noise.

However, it is difficult to estimate the underlying ground truth due to the added noise. We desire better calibration over the privacy mechanism.

E. Mechanism

The optimal mechanism for anonymized local privacy should have the following characteristics. The output space should be three values of “Yes”, “No”, and \perp . A fraction of each of the *Yes* and *No* population should be included. Each population should have some notion of plausible deniability. The total number of data owners *DO* can be computed by summing the total number of “Yes”, “No”, and \perp responses. It should be noted that if the data owners would not write \perp it would be difficult to estimate and calculate the underlying *Yes* count.

Definition 2. (Minimal Variance Parameters) We model the sum of independent Bernoulli trials that are not identically distributed, as the poisson binomial distribution (we combine “No” and \perp into the same output space for modeling purposes). Let *Yes*^l represent those that respond “Yes”, regardless if they are from the *Yes* or *No* population.

The success probabilities are due to the contributions from the *Yes* and *No* populations. We then sum and search for the minimum variance.

Utility is maximized when:

$$\min(\text{Var}(P(\text{“Yes”}|Yes)) + \text{Var}(P(\text{“Yes”}|No))) \quad (5)$$

There are a couple observations. The first is that uniform sampling across both populations (*Yes* and *No*) limits the ability to achieve optimal variance. As we increase the *No* population by increasing the queries and the number of data owners that participate, the variance will correspondingly increase. For example, 10% sampling will incur a large variance for a population of one million data owners. To address this, the sampling parameters should be separately tuned for each population. We desire a small amount of data owners to be sampled from the *Yes* population to protect privacy and an even smaller amount from the *No* population (as this population will be large and only a small amount is required for linear noise). The other observation is that for the plausible deniability, by fixing the probabilities the same across the *Yes* and *No* population also restricts the variance that can be achieved.

Thus, the optimal anonymized local privacy mechanism is one that tunes both populations.

Mechanism. Let π_s be either the sampling probability for the *Yes* population as $\pi_{s_{Yes}}$ or for the *No* population as $\pi_{s_{No}}$. Let π_p be either the plausible deniability parameter for the *Yes* population as π_1 or the *No* population π_2 respectively. The mechanism that we use which satisfies Anonymized Local Privacy is as follows:

$$\text{Privatized Value} = \begin{cases} \perp & \text{with probability } 1 - (\pi_{s_{Yes1}} + \pi_{s_{Yes2}}) \\ 1 & \text{with probability } \pi_{s_{Yes1}} \times \pi_1 \\ 1 & \text{with probability } \pi_{s_{Yes2}} \times \pi_2 \\ 0 & \text{with probability } \pi_{s_{Yes1}} \times (1 - \pi_1) + \pi_{s_{Yes2}} \times (1 - \pi_2) \end{cases} \quad (6)$$

$$\text{Privatized Value} = \begin{cases} \perp & \text{with probability } 1 - \pi_{s_{No}} \\ 1 & \text{with probability } \pi_{s_{No}} \times \pi_3 \\ 0 & \text{otherwise} \end{cases} \quad (7)$$

It should be noted that $\pi_{s_{Yes1}} \times \pi_1$ are the percentage of data owners that answer truthfully “Yes” and $\pi_{s_{Yes2}} \times \pi_2$ are the percentage of data owners that are “forced” to respond “Yes” providing the plausible deniability. Each case has its own coin toss parameters in order to be able to fine tune the variance and reduce the estimation error as opposed to the prior examples where the variance cascades across terms adding error.

V. ACCURACY

Let *DO* be the total number of data owners. Let *YES* and *NO* be the population count of those that truthfully respond “Yes” and “No” respectively such that *YES* + *NO* = *DO*.

Lemma 1. (Yes Estimate From Aggregated Count)

Expected value of “Yes” responses is:

$$\begin{aligned} E[\text{“Yes”}] &= \pi_{Yes1} \times \pi_1 \times YES + \\ &\pi_{Yes2} \times \pi_2 \times YES + \\ &\pi_{No} \times \pi_3 \times (DO - YES) \end{aligned} \quad (10)$$

Solving for *YES* results in:

$$YES = \frac{E[\text{“Yes”}] - (\pi_{No} \times \pi_3 \times DO)}{(\pi_{Yes1} \times \pi_1) + (\pi_{Yes2} \times \pi_2) - (\pi_{No} \times \pi_3)} \quad (11)$$

The estimator $Y\hat{E}S_{Yes}$ accounting for the standard deviation $\sigma(\text{“Yes”})$ is:

$$Y\hat{E}S_{Yes} = \frac{E[\text{“Yes”}] \pm \sigma(\text{“Yes”}) - (\pi_{No} \times \pi_3 \times DO)}{(\pi_{Yes1} \times \pi_1) + (\pi_{Yes2} \times \pi_2) - (\pi_{No} \times \pi_3)} \quad (12)$$

Lemma 2. Standard Deviation of the Aggregated “Yes” Count
The standard deviation $\sigma(\text{“Yes”})$ is:

$$\begin{aligned} \text{Var}(\text{“Yes”}) &= ((\pi_{Yes1} \times \pi_1 + \pi_{Yes2} \times \pi_2) \times \\ &(1 - (\pi_{Yes1} \times \pi_1 + \pi_{Yes2} \times \pi_2)) \times \\ &YES) + \\ &(\pi_{No} \times \pi_3 \times \\ &(1 - (\pi_{No} \times \pi_3)) \times \\ &NO) \end{aligned} \quad (13)$$

$$\sigma(\text{"Yes"}) = \sqrt{\text{Var}(\text{"Yes"})} \quad (14)$$

Lemma 3. (Yes Estimate From Aggregated “No” Count)

Expected value of “Yes” responses is:

$$\begin{aligned} E[\text{"Yes"}] &= \pi_{Yes_1} \times (1 - \pi_1) \times YES + \\ &\quad \pi_{Yes_2} \times (1 - \pi_2) \times YES + \\ &\quad \pi_{No} \times (1 - \pi_3) \times (DO - YES) \end{aligned} \quad (15)$$

Solving for YES results in:

$$YES = \frac{E[\text{"Yes"}] - (\pi_{No} \times (1 - \pi_3) \times DO)}{(\pi_{Yes_1} \times (1 - \pi_1)) + (\pi_{Yes_2} \times (1 - \pi_2)) - (\pi_{No} \times (1 - \pi_3))} \quad (16)$$

The estimator \hat{YES}_{No} accounting for the standard deviation $\sigma(\text{"No"})$ is:

$$\hat{YES}_{No} = \frac{E[\text{"Yes"}] \pm \sigma(\text{"Yes"}) - (\pi_{No} \times (1 - \pi_3) \times DO)}{(\pi_{Yes_1} \times (1 - \pi_1)) + (\pi_{Yes_2} \times (1 - \pi_2)) - (\pi_{No} \times (1 - \pi_3))} \quad (17)$$

Lemma 4. Standard Deviation of the Aggregated “No” Count

The standard deviation $\sigma(\text{"No"})$ is:

$$\begin{aligned} \text{Var}(\text{"Yes"}) &= ((\pi_{Yes_1} \times (1 - \pi_1) + \pi_{Yes_2} \times (1 - \pi_2)) \times \\ &\quad (1 - (\pi_{Yes_1} \times (1 - \pi_1) + \pi_{Yes_2} \times (1 - \pi_2))) \times \\ &\quad YES) + \\ &\quad (\pi_{No} \times (1 - \pi_3) \times \\ &\quad (1 - (\pi_{No} \times (1 - \pi_3)))) \times \\ &\quad NO) \end{aligned} \quad (18)$$

$$\sigma(\text{"No"}) = \sqrt{\text{Var}(\text{"No"})} \quad (19)$$

Lemma 5. (Yes Estimate From Sampled Population)

Expected value of \perp (not participating) responses is:

$$\begin{aligned} E[\perp] &= (1 - (\pi_{Yes_1} + \pi_{Yes_2})) \times YES + \\ &\quad (1 - \pi_{No}) \times (DO - YES) \end{aligned} \quad (20)$$

Solving for YES results in:

$$YES = \frac{E[\perp] - ((1 - \pi_{No}) \times DO)}{(1 - (\pi_{Yes_1} + \pi_{Yes_2})) - (1 - \pi_{No})} \quad (21)$$

The estimator \hat{YES}_{\perp} accounting for the standard deviation $\sigma(\perp)$ is:

$$\hat{YES}_{\perp} = \frac{E[\perp] \pm \sigma(\perp) - ((1 - \pi_{No}) \times DO)}{(1 - (\pi_{Yes_1} + \pi_{Yes_2})) - (1 - \pi_{No})} \quad (22)$$

Lemma 6. Standard Deviation of the Sampled Population

The standard deviation $\sigma(\perp)$ is:

$$\begin{aligned} \text{Var}(\perp) &= ((1 - (\pi_{Yes_1} + \pi_{Yes_2})) \times \\ &\quad (\pi_{Yes_1} + \pi_{Yes_2}) \times \\ &\quad YES) + \\ &\quad ((1 - \pi_{No}) \times \\ &\quad \pi_{No} \times \\ &\quad NO) \end{aligned} \quad (23)$$

$$\sigma(\perp) = \sqrt{\text{Var}(\perp)} \quad (24)$$

Lemma 7. Solving for YES

There are two approaches we can take. We can either use the “Yes” estimators to estimate the underlying population as described earlier. Or we can treat the equations as a system of linear equations.

The observation is that setting $\pi_1 = \pi_2 = \pi_3 = 1$ results in the standard deviation being equal for $\sigma(\text{"Yes"})$, $\sigma(\text{"No"})$, and $\sigma(\perp)$. This has the effect of resulting in no “No” responses and the two equations are thus dependant.

We have the following system of linear equations of two unknown variables YES and σ as follows:

$$E[\text{"Yes"}] \pm \sigma = \text{Observed}(\text{"Yes"}) \quad (25)$$

$$E[\perp] \pm \sigma = \text{Observed}(\perp) \quad (26)$$

$$E[\text{"Yes"}] \pm \sigma + E[\perp] \pm \sigma = DO \quad (27)$$

We then solve for YES and σ for each combination of varying \pm signs using a solver. We eliminate the solutions which assign YES a negative value.

It would be nice if we could cancel out the error. It would also be nice if the system of linear equations above would have exactly one solution. However, it’s not clear that we can immediately guarantee this.

VI. FIRST ATTEMPT CANCELLING THE NOISE

As we control the randomization, can we construct a mechanism whereby the error introduced by the NO population cancels out? Performing uniform sampling across both YES and NO populations allows us to cancel the *population* error though we are not able to precisely estimate the YES population as it also cancels out the YES terms.

One observation is that the error terms potentially could cancel out if the signs were flipped. Thus, we construct our mechanism as follows. Each data owner responds *twice* for the same query, though slightly flips a single term to allow for the error cancellation.

$$YES_A \text{ Privatized Value} = \begin{cases} \perp_1 & \text{with probability } \pi_{\perp_1} \\ \perp_1 & \text{with probability } \pi_Y \\ \perp_2 & \text{with probability } 1 - \pi_{\perp_1} - \pi_Y \end{cases} \quad (28)$$

$$NO_A \text{ Privatized Value} = \begin{cases} \perp_1 & \text{with probability } \pi_{\perp_1} \\ \perp_1 & \text{with probability } \pi_{\perp_N} \\ \perp_2 & \text{with probability } 1 - \pi_{\perp_1} - \pi_{\perp_N} \end{cases} \quad (29)$$

$$YES_B \text{ Privatized Value} = \begin{cases} \perp_1 & \text{with probability } \pi_{\perp_1} - \pi_Y \\ \perp_2 & \text{with probability } \pi_Y \\ \perp_2 & \text{with probability } 1 - \pi_{\perp_1} \end{cases} \quad (30)$$

$$NO_B \text{ Privatized Value} = \begin{cases} \perp_1 & \text{with probability } \pi_{\perp_1} - \pi_{\perp_N} \\ \perp_2 & \text{with probability } \pi_{\perp_N} \\ \perp_2 & \text{with probability } 1 - \pi_{\perp_1} \end{cases} \quad (31)$$

The expected values are as follows:

$$\begin{aligned} E[\perp_{1A}] &= (\pi_{\perp_1} + \pi_Y) \times YES + (\pi_{\perp_1} + \pi_N) \times NO \\ &= \pi_{\perp_1} \times YES + \pi_Y \times YES + \pi_{\perp_1} \times NO + \pi_N \times NO \\ &= \pi_{\perp_1} \times TOTAL + \pi_Y \times YES_A + \pi_N \times NO \\ &= \pi_{\perp_1} \times TOTAL + \pi_Y \times YES_A + \pi_N \times TOTAL - \pi_N \times YES \end{aligned} \quad (32)$$

$$\begin{aligned} E[\perp_{1B}] &= (\pi_{\perp_1} - \pi_Y) \times YES + (\pi_{\perp_1} - \pi_N) \times NO \\ &= \pi_{\perp_1} \times YES - \pi_Y \times YES + \pi_{\perp_1} \times NO - \pi_N \times NO \\ &= \pi_{\perp_1} \times TOTAL - \pi_Y \times YES - \pi_N \times NO \\ &= \pi_{\perp_1} \times TOTAL - \pi_Y \times YES - (\pi_N \times TOTAL - \pi_N \times YES) \\ &= \pi_{\perp_1} \times TOTAL - \pi_Y \times YES - \pi_N \times TOTAL + \pi_N \times YES \end{aligned} \quad (33)$$

$$\begin{aligned} E[\perp_{2A}] &= (1 - \pi_{\perp_2} - \pi_{\perp_Y}) \times YES + (1 - \pi_{\perp_2} - \pi_N) \times NO \\ &= (1 - \pi_{\perp_2}) \times TOTAL - \pi_Y \times YES - \pi_N \times NO \\ &= (1 - \pi_{\perp_2}) \times TOTAL - \pi_Y \times YES - (\pi_N \times TOTAL - \pi_N \times YES) \\ &= (1 - \pi_{\perp_2}) \times TOTAL - \pi_Y \times YES - \pi_N \times TOTAL + \pi_N \times YES \end{aligned} \quad (34)$$

$$\begin{aligned} E[\perp_{2B}] &= (1 - \pi_{\perp_2}) \times YES + \perp_Y \times YES + \\ &\quad (1 - \pi_{\perp_2}) \times NO + \perp_N \times NO \\ &= (1 - \pi_{\perp_2}) \times TOTAL + \pi_Y \times YES + \pi_N \times NO \\ &= (1 - \pi_{\perp_2}) \times TOTAL + \pi_Y \times YES + \\ &\quad \pi_N \times TOTAL - \pi_N \times YES \end{aligned} \quad (35)$$

We should now be able to subtract either pairs of expected values and solve for *YES*. Either $E[\perp_{1A}] - E[\perp_{1B}]$ or $E[\perp_{2A}] - E[\perp_{2B}]$. However, the variance of the total population multiplied by the π_N contributes error as the *NO* population grows.

One option is to simply create a third output where the value $\pi_N \times TOTAL$. We then can eliminate this value from both systems of equations and solve for *YES*.

VII. CANCELLING THE NOISE

The observation is that we are shifting fractions of the population across two outputs. By expanding to three outputs we can shift the population to isolate the *YES* population for estimation. We shift the *YES* and *NO* population to the third output space to blend these crowds.

$$YES_A \text{ Privatized Value} = \begin{cases} \perp_1 & \text{with probability } \pi_{\perp_1} \\ \perp_1 & \text{with probability } \pi_Y \\ \perp_2 & \text{with probability } \pi_{\perp_2} \\ \perp_3 & \text{with probability } \pi_{\perp_3} \end{cases} \quad (36)$$

$$NO_A \text{ Privatized Value} = \begin{cases} \perp_1 & \text{with probability } \pi_{\perp_1} \\ \perp_2 & \text{with probability } \pi_{\perp_N} \\ \perp_2 & \text{with probability } \pi_{\perp_2} \\ \perp_3 & \text{with probability } \pi_{\perp_3} \end{cases} \quad (37)$$

$$YES_B \text{ Privatized Value} = \begin{cases} \perp_1 & \text{with probability } \pi_{\perp_1} \\ \perp_2 & \text{with probability } \pi_{\perp_2} \\ \perp_3 & \text{with probability } \pi_{\perp_Y} \\ \perp_3 & \text{with probability } \pi_{\perp_3} \end{cases} \quad (38)$$

$$NO_B \text{ Privatized Value} = \begin{cases} \perp_1 & \text{with probability } \pi_{\perp_1} \\ \perp_2 & \text{with probability } \pi_{\perp_N} \\ \perp_3 & \text{with probability } \pi_{\perp_N} \\ \perp_3 & \text{with probability } \pi_{\perp_3} \end{cases} \quad (39)$$

The expected values are as follows:

$$\begin{aligned} E[\perp_{1A}] &= \pi_{\perp_1} \times YES + \pi_Y \times YES + \pi_{\perp_1} \times NO \\ &= \pi_{\perp_1} \times TOTAL + \pi_Y \times YES \end{aligned} \quad (40)$$

$$\begin{aligned} E[\perp_{2A}] &= \pi_{\perp_2} \times YES + (\pi_{\perp_1} + \pi_N) \times NO \\ &= \pi_{\perp_2} \times YES + \pi_{\perp_1} \times NO + \pi_N \times NO \\ &= \pi_{\perp_2} \times TOTAL + \pi_N \times NO \end{aligned} \quad (41)$$

$$\begin{aligned} E[\perp_{3A}] &= \pi_{\perp_3} \times YES + \pi_{\perp_3} \times NO \\ &= \pi_{\perp_3} \times TOTAL \end{aligned} \quad (42)$$

$$\begin{aligned} E[\perp_{1B}] &= \pi_{\perp_1} \times YES + \pi_{\perp_1} \times NO \\ &= \pi_{\perp_1} \times TOTAL \end{aligned} \quad (43)$$

$$\begin{aligned} E[\perp_{2B}] &= \pi_{\perp_2} \times YES + \pi_{\perp_2} \times NO \\ &= \pi_{\perp_2} \times TOTAL \end{aligned} \quad (44)$$

$$\begin{aligned} E[\perp_{3A}] &= (\pi_{\perp_3} + \pi_{\perp_Y}) \times YES + (\pi_{\perp_3} + \pi_{\perp_N}) \times NO \\ &= \pi_{\perp_3} \times TOTAL + \pi_{\perp_Y} \times YES + \pi_{\perp_N} \times NO \end{aligned} \quad (45)$$

We should now be able to subtract the pairs of expected values, scale the *NO* population, eliminate the error due to the population variance, and solve for *YES*. by $E[\perp_{1A}] - E[\perp_{1B}]$.

The estimation error is now only due to estimating the sampled *YES* population, so need to choose accordingly when *YES* is known to be small.

A. Privacy Algebra

To simplify reasoning regarding “shifting” the populations to protect privacy and cancel the noise due to the population variance, we introduce privacy algebra notation to simplify our expressions.

$$\begin{aligned}
 \text{Let } TOTAL &= YES + NO \\
 \text{Let } TOTAL_n &= \pi_{\perp_n} \times TOTAL \\
 \text{Let } YES_Y &= \pi_Y \times YES \\
 \text{Let } YES_f &= \pi_f \times \pi_N \times YES \\
 \text{Let } NO_f &= \pi_N \times NO \\
 \text{Let } NO_N &= \pi_f \times \pi_N \times NO
 \end{aligned} \tag{46}$$

B. Plausible Deniability

While we are able to cancel the noise and estimate the YES population, there is no plausible deniability as the YES population is exposed. We achieve this by utilizing an output space of three with three separate answers.

$$E[\perp_{1A}] = TOTAL_1 + YES_Y + NO_N \tag{47}$$

$$E[\perp_{2A}] = TOTAL_2 \tag{48}$$

$$E[\perp_{3A}] = TOTAL_3 \tag{49}$$

$$E[\perp_{1B}] = TOTAL_1 + YES_Y - YES_{f_1} - YES_{f_2} + NO_N - NO_{f_1} - NO_{f_2} \tag{50}$$

$$E[\perp_{2B}] = TOTAL_2 + YES_{f_1} + NO_{f_1} \tag{51}$$

$$E[\perp_{3B}] = TOTAL_3 + YES_{f_2} + NO_{f_2} \tag{52}$$

$$E[\perp_{1C}] = TOTAL_1 + YES_Y - YES_{f_1} - YES_{f_2} + NO_N - NO_{f_1} - NO_{f_2} \tag{53}$$

$$E[\perp_{2C}] = TOTAL_2 + YES_{f_1} + NO_{f_1} + NO_{f_{2_1}} \tag{54}$$

$$E[\perp_{3C}] = TOTAL_3 + YES_{f_2} + NO_{f_2} - NO_{f_{2_1}} \tag{55}$$

Working backwards we can start with solving for $NO_{f_{2_1}}$ to eventually solve for YES.

VIII. CONCLUSION

In this paper we demonstrate that we can add noise linear in the number of data owners to protect while preserving privacy. We introduce the family of Anonymized Local Privacy mechanisms.

REFERENCES

- [1] A. Blum, K. Ligett, and A. Roth. A learning theory approach to noninteractive database privacy. *J. ACM*, 60(2):12:1–12:25, 2013.
- [2] T. H. Chan, M. Li, E. Shi, and W. Xu. Differentially private continual monitoring of heavy hitters from distributed streams. In S. Fischer-Hübner and M. K. Wright, editors, *Privacy Enhancing Technologies - 12th International Symposium, PETS 2012, Vigo, Spain, July 11-13, 2012. Proceedings*, volume 7384 of *Lecture Notes in Computer Science*, pages 140–159. Springer, 2012.
- [3] K. Chaudhuri and N. Mishra. When random sampling preserves privacy. In C. Dwork, editor, *Advances in Cryptology - CRYPTO 2006, 26th Annual International Cryptology Conference, Santa Barbara, California, USA, August 20-24, 2006, Proceedings*, volume 4117 of *Lecture Notes in Computer Science*, pages 198–213. Springer, 2006.
- [4] J. C. Duchi, M. J. Wainwright, and M. I. Jordan. Local privacy and minimax bounds: Sharp rates for probability estimation. In C. J. C. Burges, L. Bottou, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States.*, pages 1529–1537, 2013.
- [5] C. Dwork. Differential privacy. In M. Bugliesi, B. Preneel, V. Sassone, and I. Wegener, editors, *Automata, Languages and Programming, 33rd International Colloquium, ICALP 2006, Venice, Italy, July 10-14, 2006, Proceedings, Part II*, volume 4052 of *Lecture Notes in Computer Science*, pages 1–12. Springer, 2006.
- [6] C. Dwork, K. Kenthapadi, F. McSherry, I. Mironov, and M. Naor. Our data, ourselves: Privacy via distributed noise generation. In *EUROCRYPT*, 2006.
- [7] C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In *TCC*, 2006.
- [8] C. Dwork and A. Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4):211–407, 2014.
- [9] Ú. Erlingsson, V. Pihur, and A. Korolova. RAPPOR: randomized aggregatable privacy-preserving ordinal response. In G. Ahn, M. Yung, and N. Li, editors, *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security, Scottsdale, AZ, USA, November 3-7, 2014*, pages 1054–1067. ACM, 2014.
- [10] J. A. Fox and P. E. Tracy. *Randomized response: a method for sensitive surveys*. Beverly Hills California Sage Publications, 1986.
- [11] J. Gehrke, M. Hay, E. Lui, and R. Pass. Crowd-blending privacy. In R. Safavi-Naini and R. Canetti, editors, *Advances in Cryptology - CRYPTO 2012 - 32nd Annual Cryptology Conference, Santa Barbara, CA, USA, August 19-23, 2012. Proceedings*, volume 7417 of *Lecture Notes in Computer Science*, pages 479–496. Springer, 2012.
- [12] J. Gehrke, E. Lui, and R. Pass. Towards privacy for social networks: A zero-knowledge based definition of privacy. In Y. Ishai, editor, *Theory of Cryptography - 8th Theory of Cryptography Conference, TCC 2011, Providence, RI, USA, March 28-30, 2011. Proceedings*, volume 6597 of *Lecture Notes in Computer Science*, pages 432–449. Springer, 2011.
- [13] B. G. Greenberg, A.-L. A. Abul-El, W. R. Simmons, and D. G. Horvitz. The unrelated question randomized response model: Theoretical framework. *Journal of the American Statistical Association*, 64(326):520–539, 1969.
- [14] J. Joy and M. Gerla. PAS-MC: Privacy-preserving Analytics Stream for the Mobile Cloud. *ArXiv e-prints*, Apr. 2016.
- [15] J. Joy, S. Rajwade, and M. Gerla. Participation Cost Estimation: Private Versus Non-Private Study. *ArXiv e-prints*, Apr. 2016.
- [16] S. P. Kasiviswanathan, H. K. Lee, K. Nissim, S. Raskhodnikova, and A. D. Smith. What can we learn privately? In *49th Annual IEEE Symposium on Foundations of Computer Science, FOCS 2008, October 25-28, 2008, Philadelphia, PA, USA*, pages 531–540. IEEE Computer Society, 2008.
- [17] N. Li, W. H. Qardaji, and D. Su. On sampling, anonymization, and differential privacy or, k -anonymization meets differential privacy. In H. Y. Youm and Y. Won, editors, *7th ACM Symposium on Information, Computer and Communications Security, ASIACCS '12, Seoul, Korea, May 2-4, 2012*, pages 32–33. ACM, 2012.
- [18] K. Nissim, S. Raskhodnikova, and A. D. Smith. Smooth sensitivity and sampling in private data analysis. In D. S. Johnson and U. Feige, editors, *Proceedings of the 39th Annual ACM Symposium on Theory of Computing, San Diego, California, USA, June 11-13, 2007*, pages 75–84. ACM, 2007.

- [19] R. Ostrovsky and V. Shoup. Private information storage (extended abstract). In F. T. Leighton and P. W. Shor, editors, *Proceedings of the Twenty-Ninth Annual ACM Symposium on the Theory of Computing, El Paso, Texas, USA, May 4-6, 1997*, pages 294–303. ACM, 1997.
- [20] Randomized Response. https://www.dartmouth.edu/~chance/teaching_aids/RRresponse/RRresponse.html.
- [21] A. C. Tamhane. Randomized response techniques for multiple sensitive attributes. *Journal of the American Statistical Association*, 76(376):916–923, 1981.
- [22] S. L. Warner. Randomized response: A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, 60(309):63–69, 1965.